Computer Technique for High-Speed Extraction of Speech Parameters

Mark R. Weiss and Cyril M. Harris

Reprinted from The Journal of the Acoustical Society of America, Vol. 35, No. 2, pp. 207–214, February 1963

Reprinted from The Journal of the Acoustical Society of America, Vol. 35, No. 2, 207–214, February, 1963

Computer Technique for High-Speed Extraction of Speech Parameters*

Mark R. Weiss

Federal Scientific Corporation, New York 27, New York

AND

Cyril M. Harris

Department of Electrical Engineering, Columbia University, New York 27, New York
(Received 14 November 1962)

An automatic system is described for high-speed extraction of speech parameters. Data obtained from a high-resolution, real-time spectrum analyzer of the correlation type are converted to digital format and processed by a large-scale computer. An effective program has been written in FORTRAN and FAP languages for the extraction of spectrum power, amplitudes and frequencies of the first three formants, and other functions of these parameters. Typical output data are presented.

INTRODUCTION

THIS paper describes a data reduction system that performs the following operations: it provides spectral analysis data of speech signals which are fed into the system; it processes these data at high speed by means of a digital computer; and it extracts selected parameters of speech from the analyzed data. Although the system that is described has been designed for the reduction of data from speech sounds, parameters of other acoustic phenomena can be extracted similarly by appropriate selection of system variables.

The analyzer used here is of the correlation type. In this application, a uniform frequency resolution of 63 cps across the entire frequency range of analysis is employed. Any other frequency resolution may be obtained by changing a dial setting. The Fourier analysis data are fed into an analog-to-digital conversion system that automatically writes the information on magnetic tape. This tape contains the analysis data in a digital form that is suitable for processing by an IBM 7090 computer which has been programmed to yield the desired speech parameters. In this paper, examples are given of the development of a computer program that provides the following information: formant frequencies, formant amplitudes, spectral power, time rates-of-change of these quantities, and whether the speech signal is voiced or unvoiced. Illustrations of printouts and plotouts of the data are presented.

The speech analysis and processing system that are described here have the following advantages which make feasible the processing of large volumes of speech data: (1) The spectrum analysis can be performed in real time. (2) High resolution, if desired, can be obtained. (3) Analysis data are converted to digital form automatically. (4) Many speech parameters can be extracted simultaneously in approximately five times real time. (5) The cost of data reduction per word is relatively small once a satisfactory computer program for parameter extraction is established. (6) The analyzed data are in a digital form in the computer, amenable to the study of the statistical variations of the speech parameters under investigation. (7) The Fourier analysis data that are being processed by the system can be observed by an oscilloscope while the data are being fed into the analog-to-digital system, i.e., before it goes to the computer.

* Presented, in part, at the Speech Communication Seminar, Stockholm (August 1962).

I. SPECTRUM ANALYZER

This speech analysis system makes use of a SIMORAMIC analyzer manufactured by the Federal Scientific Corporation. In effect, the output of this correlation-type device yields a spectrum analysis corresponding to that which would be obtained by use of a *continuum* of bandpass filters which have uniform bandwidth over an 8000-cps range, and which are separated in frequency by an infinitesimal amount. Figure 1 shows a block diagram of the analyzer. The speech signal to be analyzed first is heterodyned up to the region of 40 Mc, then the signal enters and circulates around a closely regulated storage loop. This unity-gain loop contains an ultrasonic-delay line which introduces a time delay T each time the signal recirculates as indicated in Fig. 1. It includes another heterodyner that acts as a time-variable phase shifter. By linear phase-modulation of the signal that is being stored (i.e., recirculated), a wave interference phenomenon takes place which leads to a Fourier analysis output after a selected number of recirculations. The frequency components in the spectrum analysis are linearly related to the time of occurrence of the appearance of these components at the output of the loop. After N recirculations, the spectrum analysis output is presented, and the contents of the loop are discharged automatically by reducing the loop gain to zero. Then a new processing period begins. Thus, spectral analyses in analog form are presented at time intervals which depend only on the number of recirculations in the loop. This number may be selected according to the analysis requirement of the speech data to be processed.

Figure 2, which is a photograph of the oscilloscopic display of the output of the analyzer, illustrates some of the characteristics of the data that it provides. The analysis is of a 24-msec segment of the vowel sound /æ/ from the spoken word "as" (/æz/). Below the main trace is a line of frequency calibration markers, spaced 500 cps apart. The harmonics of the fundamental voice frequency have an envelope with major peaks at 650, 1650, 2400, and 3800 cps. (Frequency equalization of the input speech signals is employed to enhance the higher frequencies.)

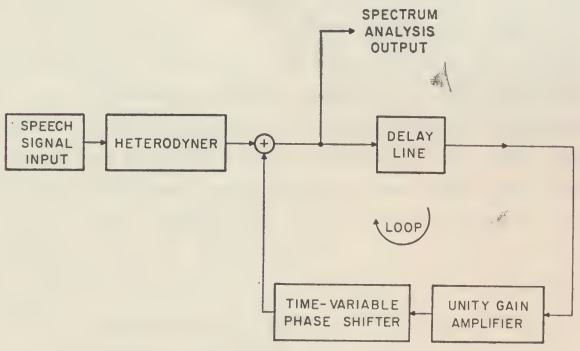

FIG. 1. Simplified block diagram illustrating the basic principles of operation of the correlation-type analyzer described here.

The resolution and frequency accuracy with which signal components can be identified in any spectrum analyzer improves directly with the signal processing time. Thus, if the processing time is 25 msec, the various frequency components can, at best, be resolved for separations no finer than 40 cps; if the processing time is 2.5 msec, the various frequency components can, at best, be resolved for separations no finer than 400 cps. In the analyzer described here, the processing time is controlled by selecting the number of recirculations N made by the speech signals in the loop of Fig. 1. During each recirculation, the signal is delayed T seconds by the delay line. It can be shown that when the number N exceeds about 30, the 3-dB bandwidth of the analyzer is approximately: $0.85/(NT)$ cps. For the system described here, the analysis time NT is 24 msec. The corresponding nominal resolution is 42 cps (i.e., 1/24 msec).

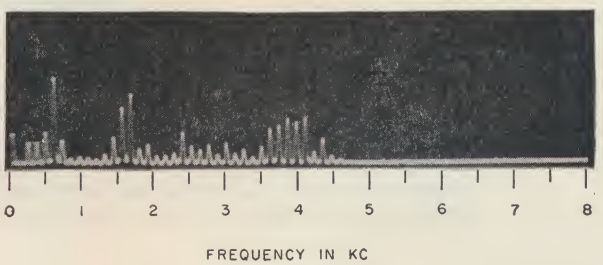

FIG. 2. Photograph of a spectrum analysis of a 24-msec segment of the vowel /æ/ with a correlation-type spectrum analyzer having a resolution of 63 cps and an analysis range of 8000 cps.

If a sine-wave signal of angular frequency ω and amplitude E is applied to the input of the analyzer, the response of the analyzer, E_{out}, is given by

$$E_{\text{out}} = E \frac{\sin[\frac{1}{2}(N+1)(\omega T - \omega_d \tau)]}{\sin[\frac{1}{2}(\omega T - \omega_d \tau)]},$$

where ω_d is the angular frequency of the time-variable phase shifter, and where $0 \leq \tau < T$. Thus, the response function of the analyzer is of the form $(\sin Nx)/(\sin x)$, which is shown as the dashed line in Fig. 3. Since this analyzer is of the type which operates in the time domain, the shape of this output response characteristic can be modified easily by means of amplitude modulation (i.e., "weighting") of the analyzed signal during the signal processing period. For the work described here, triangular weighting is employed, i.e., the weighting factor is zero initially, increases linearly to a maximum value of unity midway in the processing period, and then falls to zero again in a symmetrical

fashion. The "side lobes" in the frequency response characteristic are reduced greatly in exchange for a broadening of the 3-dB width of the response peak, and a broadening of the nominal resolution from 42 to 63 cps. This is illustrated by a comparison of the "weighted" response, shown by the solid line in Fig. 3, with the unweighted response shown by the dashed line. This application of amplitude weighting in the time domain is analogous to the amplitude weighting in linear arrays of hydrophones or antennas. Whereas, uniform weighting provides optimum processing for signals accompanied by a significant amount of noise or which have frequency components that are relatively closely spaced in frequency, it is more desirable to use the triangular weighting here because of the reduction in "side-lobe" response.

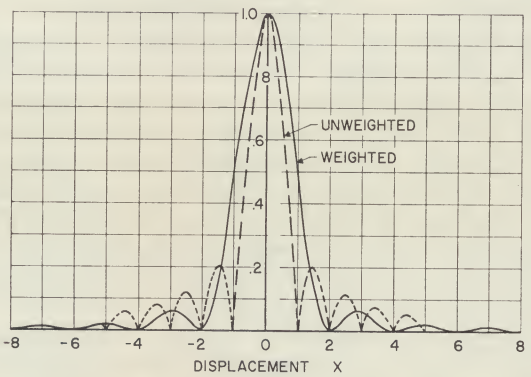

FIG. 3. Frequency selectivity characteristics of the analyzer for "unweighted" and "triangular weighted" input signals. The abscissa represents displacement from the signal frequency, measured in terms of a dimensionless quantity X. For any value of processing time T the displacement in cycles per second is obtained as X/T, where T is measured in seconds.

II. ANALOG-TO-DIGITAL READOUT SYSTEM

The analyzer described above provides a sequence of spectral data, in the form shown in Fig. 2. In order to process these data by means of a digital computer that has the capacity for handling information at a high bit rate, the analog-to-digital readout system illustrated in Fig. 4 has been constructed. Since the resolution of the data which are supplied by the analyzer is high, the sampling rate must be sufficiently high, and the scanning aperture must be sufficiently narrow, to avoid a loss of useful spectral information. The samples of the speech spectrum are converted to six-bit binary numbers by the analog-to-digital converter. The converter is capable of providing seven-bit accuracy at sampling rates up to 3.3 Mc with a deviation from linearity of less than half a bit.

Although the IBM data-processing equipment which is used in this system can process in real time the

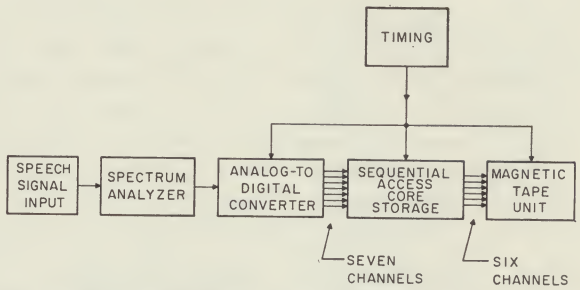

FIG. 4. Block diagram of the analysis and conversion system used in the generation of digital recordings of speech spectra.

average output supplied by the analyzer, a storage unit is required to handle the peak bursts of information which are provided after each 24 msec of signal processing. This sequential-access core-memory "storage-and-hold" unit supplies the converter output information to the magnetic tape unit at a 42-kc rate. The memory is capable of storing 256 seven-bit words. The information stored in the magnetic core memory is read out in approximately 6 msec, and is fed into an IBM 729-II magnetic tape unit. This device is a digital tape recorder operated in the high-density mode, and used in conjunction with the IBM 7090 computer.

At the present time each Fourier analysis in analog form is sampled at a rate of 330 kc by the analog-to-digital converter. This limitation on the sampling rate is imposed by the storage rate capability of the magnetic core-memory unit. By use of a higher performance core unit, or by use of several of the type now being employed, the sampling rate can be greatly increased. The rate currently in use corresponds to a sampling interval of 30 cps in the frequency domain across the entire frequency spectrum—a rate high enough to ensure sampling within 1 dB of any peak of a pitch harmonic. By use of the estimation procedure described in the next section of this paper, the frequency of any pitch harmonic can be determined with an accuracy of 4 cps. Higher accuracy, if required, can be achieved by increasing the density of the samples, i.e., by reducing the separation between samples in the frequency domain.

III. COMPUTER PROGRAM

The first step in the development of the computer program described here was to obtain a sequence of photographs, with a motion picture camera, showing a sequence of Fourier analyses as a function of time. These photographs are of the type shown in Fig. 2. Then, a program was written describing the sequence of mechanical steps to be taken in reducing the photographed Fourier analysis data to provide the required speech parameters as a function of time. These steps are referred to here as a "mechanical program." The

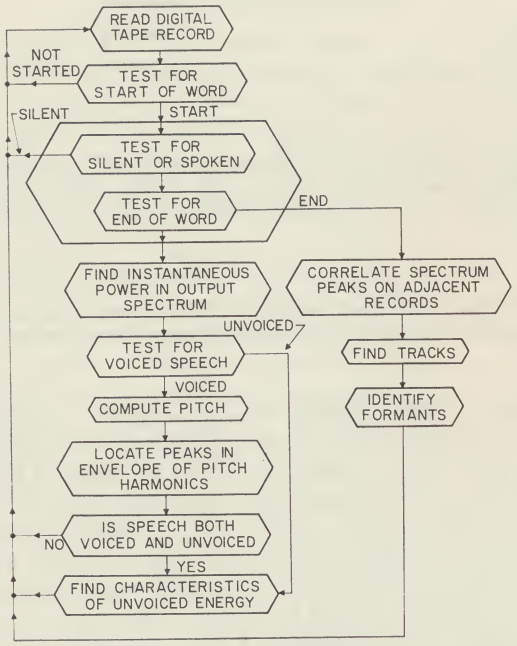

FIG. 5. Sequence of steps taken in the computer program for the extraction of speech parameters from digital recordings of speech spectra.

photographed analog information then was manually converted to digital information, and the digital information so obtained was processed with a desk calculator. The mechanical program which finally was developed provided both a statement of the problem to be solved by the digital computer and the solutions for specific examples, giving the values one should obtain for the extracted speech parameters as a function of time. These carefully solved examples provided a check on the results first obtained by the digital computer—a vital step in the "debugging" process of the computer program development.

This first step is an essential part of the development of the program, since usually it is found that the first statement of the problem (i.e., a list of the steps to be taken in reducing the data) does not yield the extracted parameters with the desired accuracy or in an optimum manner. For example, in the development of the present program, many data-reduction routines were worked out for formant extraction. Then, in order to improve the accuracy of the formant extraction, a correlation routine was tested in the manual program. In this routine, greater weight was given to the extracted formant frequency if there were a high degree of correlation between this value and the values obtained in the preceding and succeeding processing periods. A manual procedure, which included a correlation routine, was developed which finally produced satisfactory results. This procedure then was included in the computer program. Again, manual data were available for checking the results of the computer output. The second phase of the program consisted of translating the mechanical program into an equivalent computer program written in FORTRAN and FAP languages.

The logic flow of the present program is illustrated in Fig. 5. Details of the steps that lead to the extraction of the desired parameters are as follows:

(*1*) *Read digital tape record.* A digital tape record is defined here as a magnetic recording, in proper format, which represents the spectrum analysis of a single processing period of the input signal. All of the records for one word are read into the computer in sequence. The data contained in each record represent the amplitudes of the spectrum at each of the 256 sample points spaced at 30-cps intervals across the entire frequency range. This information is stored in 256 successive locations of the memory of the computer.

(*2*) *Test for "start of word."* The number of samples that exceeds a preset threshold is counted. If this number is greater than 5, a "start of word" decision is made at this point; if the number is between 3 and 5, the next digital record is examined; and if the number in this record exceeds 5, the preceding record is identified as the start of a word. If no "start of word" decision is made for the record under examination, then the search continues with the examination of successive records until such a decision is made.

(*3a*) *Test for "silent" or "spoken."* The decision "silent" is made if no samples exceed the threshold in step 2. If the decision "silent" is made, the computer program continues with the reading of the next record. If the threshold level is exceeded, the decision "spoken" is made and the program proceeds to the next step.

(*3b*) *Test for "end of word."* If the threshold level is not exceeded four times in succession, an "end of the word" decision is made, and the program proceeds with step 10.

(*4*) *Compute instantaneous power in spectrum.* The instantaneous power in the output spectrum is obtained by squaring and summing the amplitudes of the stored samples during each record. (Account is taken of the equalization characteristic employed during the recording of the words which are analyzed—flat up to 1000 cps, then rising at the rate of 6 dB/octave to 8000 cps.)

(*5*) *Test for "voiced" or "unvoiced" speech.* In the test for voicing, the ratio of the power in the spectrum below 1000 cps to the total power in the spectrum is determined. If this ratio does not exceed a reference value, the decision "unvoiced" is made and the program continues with step 9. If the reference value is exceeded, the decision "voiced" is made.

(*6*) *Compute pitch.* Pitch is computed by finding the average spacing between n pitch harmonics in a se-

lected group of pitch harmonics. (A detailed description of the pitch extraction procedure is to be presented by the authors in a paper in the March issue of the Journal.)

(*7*) *Locate the peaks in the envelope of pitch harmonics.* The formants of a speech sound in any record appear as peaks in the envelope of pitch harmonics in its spectrum. All peaks are located for each individual record. In the previous step, those samples of the speech spectrum which were closest to the frequencies of the pitch harmonics were identified as the "pitch samples." Here, these pitch samples are examined; those which are both larger than or equal to the succeeding pitch sample and larger than the preceding one are noted. They represent the locations of the peaks in the envelope of pitch harmonics within close tolerance.

(*8*) *Test to determine if speech contains both voiced and unvoiced energy.* The ratio of the power in the speech spectrum above 5000 cps to the total power is computed. If this ratio is below a threshold level, the decision is made that the speech sound contains only voiced energy, and the program continues with step 1 for the next digital record. Otherwise, a decision is made that the speech may contain both voiced and unvoiced energy. Next, the frequency ranges of unvoiced speech energy above 5000 cps are identified. If the spacing between sample peaks in any range above 5000 cps is regular and equal to the spacing between pitch samples, then the energy in this portion of the spectrum is associated with voicing. If the spacing is not regular, the frequency range under examination is identified as corresponding to an unvoiced speech sound; this results in the decision that the sound contains both voiced and unvoiced energy.

(*9*) *Compute characteristics of unvoiced energy.* The limits of each frequency range of unvoiced energy are identified. Five successive zero-amplitude samples constitute this limit. The average power spectral density in 30-cps bands within each frequency range is computed.

Each digital record within a speech word is processed in accordance with the preceding steps. When this information has been obtained for all records in the word, it is used to locate formants as follows:

(*10*) *Correlate spectrum peaks on adjacent records.* The amplitudes and locations of the peaks which are identified as possible formants (determined for each digital record in step 7) are compared with similar data on adjacent records. The purpose of this step is to reject any spurious peaks and to increase the accuracy of tracking of the peaks. For each pair of peaks on adjacent records being compared, the sum of the squares of the differences between the amplitudes and frequencies is determined. A peak on any one record is considered to be related to that peak on an adjacent record for which the sum of the squares of the differences is least.

(*11*) *Find tracks.* A tracking procedure is initiated with the first "voiced" record of the word being analyzed, and proceeds through all the succeeding voiced records. A "track" is defined here as a series of envelope peaks appearing on successive records, which are highly correlated from record to record. A track of less than three related peaks is considered spurious.

(*12*) *Identify formants.* The tracks which are determined in the preceding step include *F*1, *F*2, *F*3, and minor tracks which occur both below, between, and above these formants. To aid in the identification of the formants, each of the tracks located in step 11 is described in terms of the following characteristics: "total power" represented in the track; the average, minimum, and maximum frequencies of the track; and the time duration of the track. In the identification procedure, the following constraints are imposed on the frequency ranges for the formants: *F*1 (250 to 1000 cps), *F*2 (500 to 2800 cps), and *F*3 (1200 to 3300 cps). Both the total power and the "average" frequency of the track are taken into account in identifying the formants in each range. The use of these data as discriminants for the formants takes into account the overlap of the formant ranges, and also allows for the possible occurrence of a greater sound energy in *F*3 than in *F*2.

IV. EXAMPLES AND CONCLUSIONS

The output of the processing system is presented in several ways which have been found useful in our research program. In general, it is not economical to use the readout system of the IBM 7090 computer directly. Instead, after processing, the output data (in the form of digital tape recordings) are fed into an IBM 1401 computer and its associated printout equipment. Table I shows a typical numerical printout so obtained. The numerals listed under "Frame" in column 1 correspond to the number of processing periods that have elapsed since the start of the word. For example, "Frame 1" data represent the data averaged over the first 24 msec. The second group of columns shows the spectrum power obtained by squaring and summing the amplitudes of the spectrum samples; these data are presented in terms of their linear and logarithmic values, together with their time derivatives. In each of the three remaining groups of columns are listed: the levels of the formants, formant frequencies, logarithms of the formant frequencies, and rates of change of frequency.

Often a significant correlation between parameters can be observed visually if the data are presented in graphical form. Therefore, in addition to the precise numerical values, a computer plotout is obtained from the IBM 1401 computer. For example, the data shown in tabular form in Table I are shown, as automatically

and "slow" speech were time-normalized to provide an over-all length that is the same as that for "normal" speech. Comparisons of this type require detailed characteristics of the parameters of speech throughout the phonation of the words. When these parameters are extracted, as a function of time, for a large number of words, considerable data reduction is involved. A system of the type described makes practical the acquisition of such large quantities of reduced data. Furthermore, the data that are obtained are in a form such that it is convenient to test various hypotheses, concerning the parameters so extracted, by the use of suitable computer programs.

ACKNOWLEDGMENTS

This work was supported by the Rome Air Development Center, U. S. Air Force. The authors are grateful for the kind cooperation of Dr. Gordon E. Peterson under whose supervision recordings were made of the words used in testing this automatic data-reduction system.

LANCASTER PRESS, INC., LANCASTER, PA.

Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 36, NUMBER 5 MAY 1964

Effects of Speaking Condition on Pitch*

CYRIL M. HARRIS

Department of Electrical Engineering, Columbia University, New York, New York 10027

AND

MARK R. WEISS

Federal Scientific Corporation, New York, New York 10027

(Received 30 January 1964)

Pitch changes that occur when a talker alters the speaking condition of his voice from "normal" to "loud" or to "soft," for otherwise natural speaking conditions, were evaluated for two groups of recordings: (1) recordings of the same word, each spoken by a different talker having "general American" dialect, and (2) recordings of a group of different words spoken by the same talker—a trained phonetician. Using normal speech as a basis of comparison, the results of the first group of words for 14 talkers shows an average increase in pitch of 34% for loud speech and a decrease in pitch of 12% for soft speech. Similarly, the results for the one talker saying 28 words shows an average increase in pitch of 53% for loud speech and a decrease of 11% for soft speech. A third group of recordings of speech spoken in a monotone was analyzed to determine the standard deviation of the pitch fluctuations about the mean value of pitch for each word. For a group of 21 untrained talkers the average fluctuation in pitch was 4.4%, in contrast with a value of 2.4% for a trained phonetician. Other results showed that an increase in speech-power level from normal to loud conditions resulted in an increase in the frequency of formant 1 of 11%, but no significant shift in formants 2 or 3 was observed.

INTRODUCTION

THE qualitative observation that the pitch of one's voice is dependent on the type of speaking condition has been noted by many research workers in the past. In this paper, statistical data are presented concerning the effects of speaking condition on pitch. Statistical data are given of the shift in pitch that occurs when a talker changes his voice from "normal" to "loud" or to "soft" under otherwise natural speaking conditions. For the case of monotone speech, results are given in terms of the standard deviation of the pitch fluctuations about the mean value of pitch for each word. In addition, data are given of the shift in frequency of the first formant (F1) for an increase in speech-power level.

The speech-analysis system that was used for obtaining the data presented here has been described in detail in previous papers.[1,2] This system extracts and prints out a number of speech parameters as a function of time—for example, formant frequencies and amplitudes. These parameters are computed automatically from real-time high-resolution spectral analyses of speech. This information is converted to digital form, suitable for processing on the IBM-7090 computer. In essence, pitch versus time is obtained from a computer subroutine by determining the average frequency difference between a number of successive pitch harmonics.[3] In this computer subroutine, groups of successive pitch harmonics are examined to determine which of the groups will provide the most-accurate results. From the selected group, the value of pitch is computed to an accuracy of about 2 cps. This process is carried out at 24-msec intervals. Formant-frequency

* Presented, in part, at the Sixty-Sixth Meeting of the Acoustical Society of America, 6–9 November 1963.

[1] M. R. Weiss and C. M. Harris, J. Acoust. Soc. Am. **35**, 207–215 (1963).

[2] M. R. Weiss and C. M. Harris, J. Audio Eng. Soc. **12**, 147 (1964).

[3] C. M. Harris and M. R. Weiss, J. Acoust. Soc. Am. **35**, 339–(1963).

data are obtained from the high-resolution spectrum analyses by a computer program that locates the peaks in the envelope of the individual pitch harmonics.

I. RECORDING THE TEST WORDS

The data presented here were extracted from two groups of recordings: (1) the same word spoken by a number of different talkers for different conditions of speech, and (2) different words spoken by the same talker for different conditions of speech. These words were selected from a list of 30 words having the highest relative frequency of occurrence for written English.[4] This source material was recorded by adult male talkers classified as having a "general American" dialect. Four of the group of 21 talkers had had some phonetic training.

The words selected for analysis (i.e., the test words) were recorded for the different conditions of speech by each of the talkers. These speech recordings were made in a free-field room at the Communication Sciences Laboratory of The University of Michigan, under the supervision of Prof. Gordon E. Peterson and under the immediate control of June Shoup.

Each talker was given a set of printed instructions describing the condition of speaking desired for each recording of the list of test words. For the normal condition, there were no specifications as regards intonation or stress—the talkers were simply asked to speak as naturally as possible; for the loud condition, the talkers were asked to speak in as strong or loud a voice as possible without straining or shouting; for the soft condition, the talkers were asked to speak in as weak or soft a voice as possible with the provision that the words be voiced (not whispered). For each of these conditions the talkers maintained a normal word duration; that is the words were neither stretched nor spoken rapidly. The talkers rehearsed the list of test words under the specified condition of speaking prior to each recording of the list. Only if a talker failed to observe the instructions was there any correction of his manner of speaking. In order to obtain recordings representative of natural conditions, no attempt was made to match vowel sounds or otherwise alter the manner in which the talkers spoke when observing these instructions.

The response of the electroacoustic system employed in making the recordings of the words did not vary by more than ± 1 dB throughout the frequency range from 100–8000 cps. The recordings were made on an Ampex tape recorder, model 300-4. The combined characteristics of the recorder and magnetic tape provided a frequency-response characteristic that was essentially flat over this frequency range, with dynamic range in excess of 50 dB. The recording level was maintained as close as possible to -6 dB on the VU meter of the recorder. The following equalization characteristic was used in recording: constant from 0–1000 cps; then rising at the rate of 6 dB per octave up to 8000 cps; no additional emphasis is provided. Calibration signals, recorded at the same time as the speech material, checked the over-all operation of the system.

II. RESULTS

The results presented in this section are based upon pitch data of the type shown in Fig. 1(a). This illustration shows pitch as a function of time for a single utterance of the word *as* (/æ z/) for normal and loud conditions of speech for one talker. The data points, obtained at 24-msec intervals, have been joined by straight lines to indicate the pitch tracks.

In order to present statistical results for pitch shifts for a number of different utterances, some form of time-normalization is required, since the pitch tracks will not be of the same length. Such a normalization is accomplished by replotting the data, employing "length of pitch track, in percent" as the abscissa—0% being the *start of the word*, 100% being the *end of the word*. The pitch data of Fig. 1(a) are shown in this form in Fig. 1(b). For convenience, the normalized data are computed (from the data obtained at 24-msec intervals) at five points along the pitch track; these points (at 0%, 25%, 50%, 75%, and 100%) are joined by straight lines. The shifts in pitch that occur in changing from normal to loud speech also can be expressed in terms of a *percentage change in pitch*, relative to normal speech. The data of Fig. 1(b) are replotted in this manner in Fig. 1(c). It is in this more convenient form that the following results are presented.

A. Variation in Pitch with Power Level

The data presented in this section show how pitch varies with the power level of natural speech for three power levels: loud, normal, and soft. There was considerable variation in level during the recording of the test words for the loud condition, depending on the speaker, the word, and his particular utterance of the word. However, the loud condition, on the average, is estimated to be approximately 10 dB above normal speaking level, and the soft condition approximately 10 dB below normal speaking level.

1. Same Word; Different Talkers

A group of 14 different talkers were asked to say the same word *the* (/ðə/) three different times under natural speaking conditions. For each utterance, the pitch was determined at 24-msec intervals. Then, the percent variability in the average pitch was computed for the same word spoken by the same talker.[5] This value, averaged for the group of speakers is 3.3%. The standard deviation about this average value is 2.3%.

[4] H. Fletcher, *Speech and Hearing in Communication* (D. Van Nostrand Co., Inc., Princeton, N. J., 1953), Table 8, p. 90, after G. Dewey.

[5] This value represents the variability in average pitch for a given word when there are 29 other spoken words between repeated utterances.

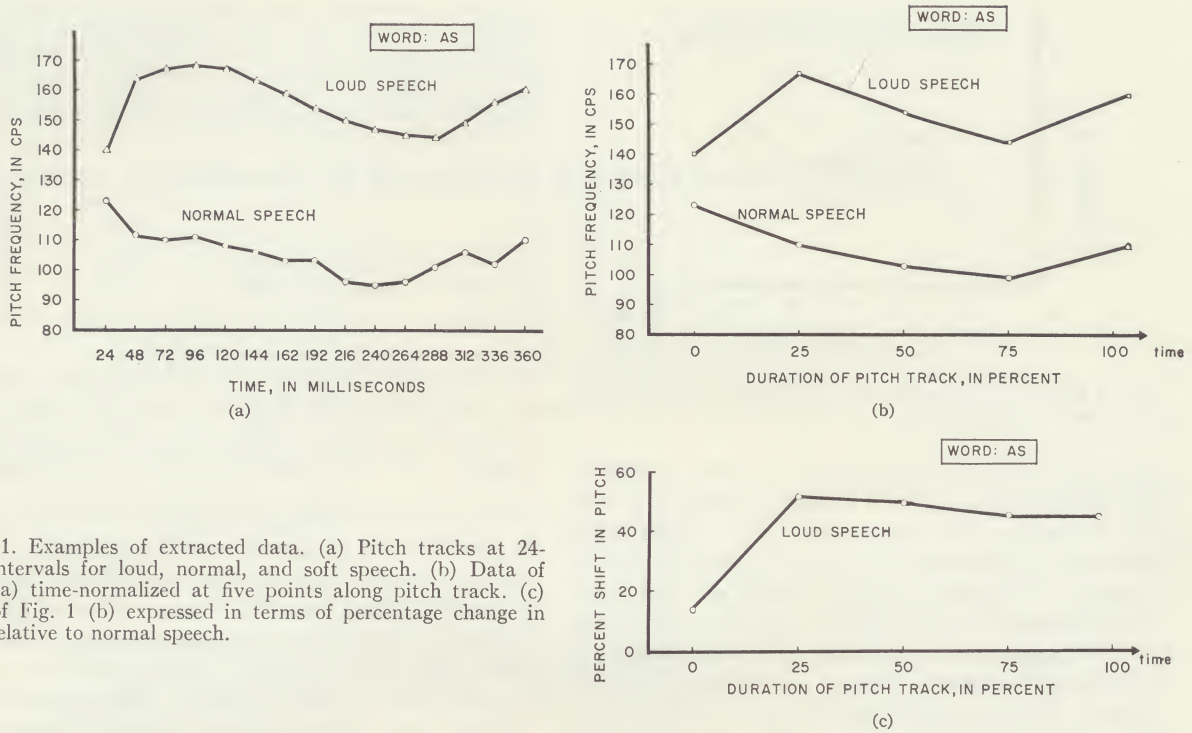

FIG. 1. Examples of extracted data. (a) Pitch tracks at 24-msec intervals for loud, normal, and soft speech. (b) Data of Fig. 1(a) time-normalized at five points along pitch track. (c) Data of Fig. 1 (b) expressed in terms of percentage change in pitch relative to normal speech.

The percentage increase in pitch that occurs in going from normal speech to loud speech is shown in Fig. 2 as a function of time by the upper solid curve. Each point on the curve represents the average increase in pitch at that value of the abscissa for the entire group. If these results are averaged over time, an average increase in pitch of 34% is obtained for loud speech. In a similar manner, the lower solid curve shows the decrease in pitch that occurs when changing from normal speech to soft speech. The decrease in pitch in this case, when averaged over time, is 12%. For both loud and soft conditions or speech, the standard deviation σ, shown as a function of time, is represented by dashed lines. Note that the standard deviation increases at the end of the pitch shift curves in both cases.

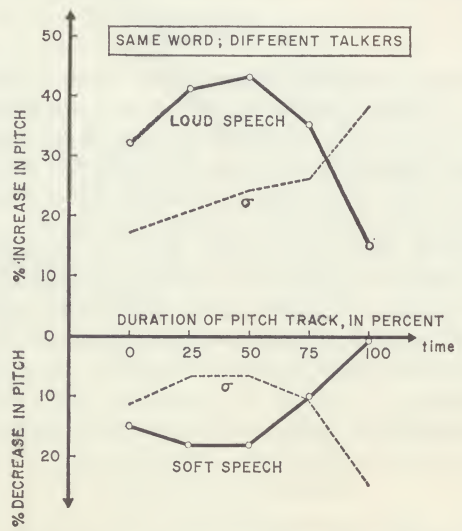

FIG. 2. Shifts in pitch for loud and soft speech observed for different talkers saying the same word.

2. *Same Speaker; Different Words*

Figure 3 summarizes the results averaged over a group of 28 different words, all spoken by the same talker (GEP). This speaker is a research phonetician who has had considerable experience in the field. The

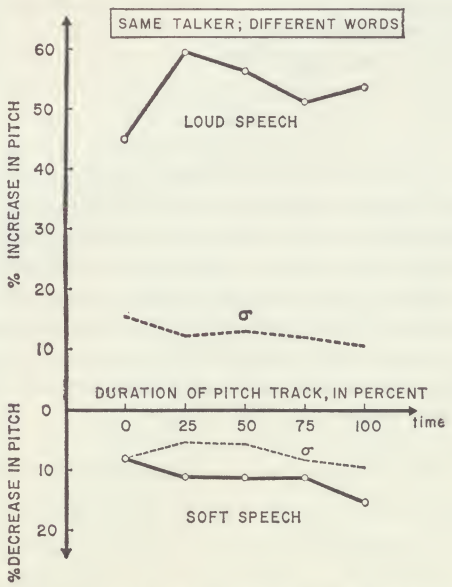

FIG. 3. Shifts in pitch for loud and soft speech observed for the same talker saying many different words.

Implementation of a Pitch Extractor of the Double-Spectrum-Analysis Type

Mark R. Weiss, Reinhold P. Vogel, and Cyril M. Harris

Reprinted from The Journal of the Acoustical Society of America, Vol. 40, No. 3, pp. 657–662, September 1966

Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 40, No. 3, 657–662, September 1966

Implementation of a Pitch Extractor of the Double-Spectrum-Analysis Type

MARK R. WEISS AND REINHOLD P. VOGEL

Federal Scientific Corporation, New York, New York 10027

CYRIL M. HARRIS

Department of Electrical Engineering, Columbia University, New York, New York 10027

A new technique of pitch extraction has been implemented that makes use of the double spectrum analysis of speech signals. In contrast to previous applications of this method of extracting pitch, the system described here operates in real time, it does not require the use of a computer, important system parameters such as "processing period" can be changed manually or adaptively, and the use of logarithmic compression is optional. When logarithmic compression is employed, the pitch extractor represents a real-time implementation of the "cepstrum" technique. The performance characteristics are described and illustrated for practical field conditions, including poor signal-to-noise ratios.

INTRODUCTION

DURING the past several decades, considerable engineering effort has been devoted to the development of a reliable pitch-extraction device for use in providing pitch information for speech bandwidth-compression systems. Such information includes both a determination of whether speech is voiced or unvoiced and—if voiced—a measurement of the fundamental pitch frequency. Most of the pitch-extraction devices that have been developed require the presence of a significant fundamental component in the speech signal, and are particularly susceptible to error if the input-speech signal contains noise, hum, or extraneous sinusoidal signals. Consequently, under practical conditions, they often fail to provide reliable pitch data. In recent years, several investigators have described an approach that avoids these shortcomings. In 1962, Bogert, Healy, and Tukey introduced a new concept, using the power-spectrum analysis of the logarithm of the power spectrum for analyzing signals.[1] They called this technique "cepstrum" analysis. A computer simulation of the process was employed to analyze seismic waveforms. Weiss and Harris,[2,3] in 1962 and 1963, described a pitch-analysis technique in which the spectra of speech signals, obtained by use of a real-time spectrum analyzer, were analyzed in a computer to extract pitch by determining the frequency difference between successive pitch harmonics. They reported in 1963 (Ref. 4) and 1964 (Ref. 5) on the use of this system in a study of variations in vocal pitch with levels of speech. In 1964, Noll and Schroeder[6,7] modified the cepstrum-analysis technique of Bogert *et al.*, defining a short-time cepstrum as the short-time power spectrum of the logarithm of the short-time power (or amplitude) spectrum. They demonstrated a computer-simulated vocoder, using pitch excitation signals obtained by computer simulation of their cepstrum-analysis technique.

The system whose implementation is described in this paper differs from the above techniques in the following ways: (1) it does not require the use of a computer;

[1] R. P. Bogert, M. J. Healy, and J. W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking," presented at the Symposium on Time Series Analysis, Brown Univ., June 1962; in *Proceedings of the Symposium on Time Series Analysis*, M. Rosenblatt, Ed. (John Wiley & Sons, Inc., New York, 1963), pp. 209–243.

[2] M. R. Weiss and C. M. Harris, "High-Resolution Analysis of Speech by Computer Techniques," in "Abstracts of Papers on Speech Analysis, Stockholm Speech Communications Seminar, 1962, Royal Institute of Technology, Stockholm, Sweden," J. Acoust. Soc. Am. **35**, 1112 (A) (1963).

[3] C. M. Harris and M. R. Weiss, J. Acoust. Soc. Am. **35**, 339–343 (1963).

[4] C. M. Harris and M. R. Weiss, J. Acoust. Soc. Am. **35**, 1876(A) (1963).

[5] C. M. Harris and M. R. Weiss, J. Acoust. Soc. Am. **36**, 933–936 (1964).

[6] A. M. Noll, J. Acoust. Soc. Am. **36**, 296–302 (1964).

[7] A. M. Noll and M. R. Schroeder, J. Acoust. Soc. Am. **36** 1030(A) (1964).

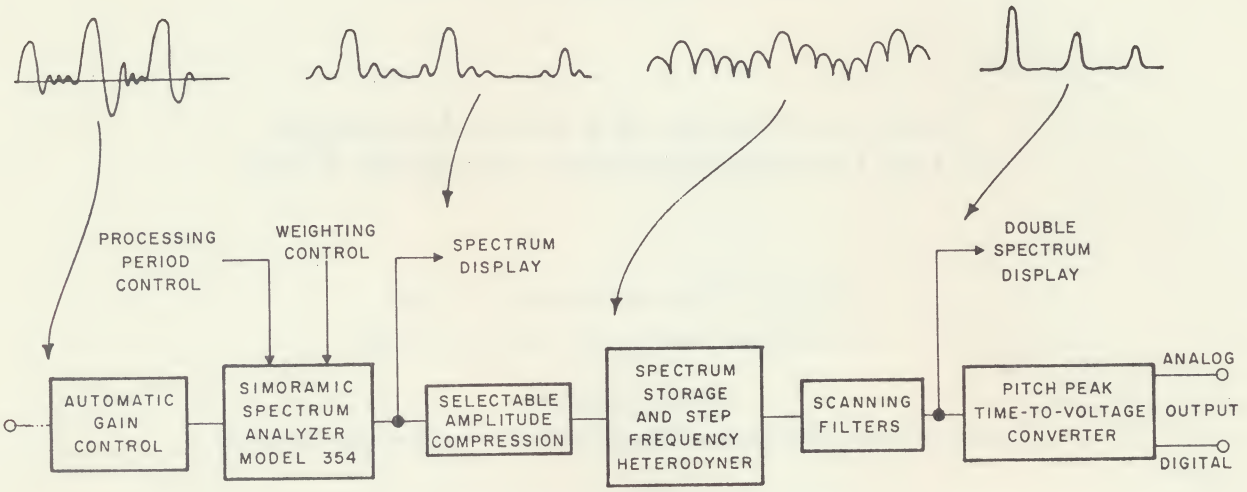

FIG. 1. Simplified block diagram of pitch extractor.

(2) it operates entirely in real time; (3) the use of logarithmic compression is optional; and (4) important system parameters such as "processing period" can be changed manually or adaptively. This flexibility makes possible various modes of analyses. For example, when logarithmic compression is employed, the pitch extractor represents a real-time implementation of the cepstrum technique.[1,6,7] When logarithmic compression is not employed, the pitch extractor determines the short-time power spectrum of the short-time amplitude spectrum.

Preliminary results show that the pitch extractor has the following performance characteristics in the pitch range from 50 to 500 cps:

- Pitch-measurement accuracy is better than 2% in the signal-to-noise (S/N) range from +30 to +10 dB, and voiced/unvoiced classifications are made with virtually 100% accuracy. Under these conditions, logarithmic compression is advantageous and the double-spectrum output of the pitch extractor is equivalent to a cepstrum.

- At about 0-dB S/N, pitch extraction becomes slightly noisy, but is still very good (the accuracy is between 3% and 4%). At this S/N ratio, or for poorer conditions, logarithmic compression is not advantageous.

- At about −10 dB S/N (speech is substantially unintelligible in this condition), pitch extraction is further degraded, although useful extraction of pitch values is possible. Voiced/unvoiced decisions can still be made by the pitch extractor. Reliable pitch extraction under unfavorable S/N ratios is possible with this technique because use is made of the energy in all of the significant harmonic components of voiced speech sounds within the analysis band.

- Elimination of speech frequencies up to 300 cps does not significantly affect pitch extraction; elimination of speech frequencies up to 500 cps has but a minor effect on pitch-extractor performance. Thus, pitch can be extracted even if the fundamental and many pitch harmonics are missing from the spectrum waveform.

- The extraction technique is independent of the relative phases of the pitch-harmonic components of the input-speech signal.

- A speech signal is classified as "voiced" only when the spectrum of the signal exhibits a pattern of regularly spaced peaks, a condition that is satisfied only when the signal contains a number of harmonically related components, so that spurious input signals such as random noise or sine waves (60 cps, for example) will not result in "voiced" speech decisions.

I. DOUBLE-SPECTRUM PITCH EXTRACTOR

A. General Description

A basic block diagram of the double-spectrum pitch-extractor is given in Fig. 1, and its sequence of operations is shown in Fig. 2. The speech signal is applied to the input of the first spectrum analyzer, which determines the spectral content of the signal during successive time intervals, i.e., "processing periods" of the analyzer. At the end of each processing period, of duration P sec, the spectrum waveform, such as shown in Fig. 1, is displayed and simultaneously fed through an amplitude compressor to a storage device where it is retained. The storage device is a wide-band, unity-gain loop that contains an ultrasonic delay line and a heterodyner. The compressed spectrum waveform is stored in this loop and circulated without distortion. After transfer of the spectrum waveform, the contents of the analyzer are discharged completely, and a new processing period is begun.

During the new processing period, the spectrum of the stored waveform is scanned in small frequency steps by a pair of filters that are part of the second spectrum analyzer in the system. Frequency stepping of the wave-

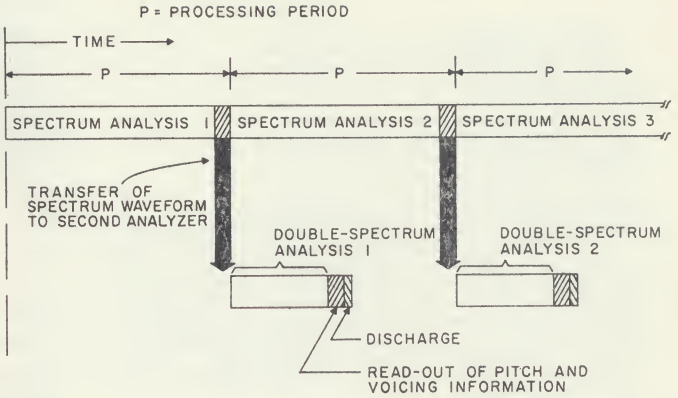

FIG. 2. Timing sequence in double-spectrum analysis cycle.

form is achieved by the heterodyner in the storage loop.

If the highest peak in the output of the second spectrum analyzer (referred to as the "double-spectrum waveform") exceeds a selected threshold value, a decision is made that the input signal contains "voiced" speech, and the elapsed time between the start of the waveform and the peak is measured. This time interval, which is proportional to the pitch period, is expressed both as an analog and a digital signal of the output of the pitch extractor.

B. Design Considerations

The successful implementation of the double-spectrum technique places severe requirements on the spectrum analyzers. Since this pitch extraction technique is based upon the measurement of frequency differneces between successive pitch harmonics, the peaks in the spectrum waveform that represent pitch harmonics must be located accurately, and their separation must be linearly proportional to pitch frequency. Thus, in order for the pitch extractor described here to achieve the maximum potential performance, the first spectrum analyzer should provide an analysis that is close to ideal. Several types of spectrum analyzers were considered for use in obtaining the first spectrum. (1) *Heterodyne Analyzer*—It is possible to determine the location of the peaks on the spectrum by the use of a swept-frequency scanning filter, but this type of frequency analyzer operates in nonreal time. (2) *Bank of Filters Analyzer*—A spectrum analyzer consisting of a bank of filters may be used in real time. However, in order to determine the exact location of the peak of the spectral components in this type of analyzer, interpolation between filters usually is required because the analyzer consists of a limited number of filters. Consequently, there is a loss of accuracy in the determination of pitch. (3) *Time-Domain Analyzer*—In the implementation described here, a model 254 SIMORAMIC® spectrum analyzer[8] was employed because it achieves a very nearly ideal representation of the Fourier transform of a speech-input signal in real time. By means of a time-varying network in a feedback system, this instrument generates the spectrum waveform on a continuous and linear frequency scale[9] and it has a linear dynamic range in excess of 45 dB. In effect, the spectrum analyzer synthesizes a continuum of infinitesimally spaced filters of uniform bandwidth in the frequency range from 25 to 4000 cps. Thus, in the waveform produced at the analyzer's output, each spectral component results in a peak that occurs exactly where it should on the spectrum frequency scale and the waveform is presented as an electrical signal in which frequency is linearly transformed into time.

The major peak on the double-spectrum waveform must be located with the same accuracy as are the harmonic peaks on the signal-spectrum waveform. Since, in this case, only one peak need be located, it is not necessary to employ as fine a quantization as is provided by the SIMORAMIC analyzer in performing the second spectrum analysis. Instead, a stepped-frequency scanning filter type of spectrum analyzer can be used here. The required accuracy in locating the major peak on the double-spectrum waveform can be achieved by use of "pulse-splitting" techniques (similar to those used in some radar systems). In determining the pitch period at the output of the second spectrum analyzer, it is convenient to present the frequency components of the spectrum waveform on a linear scale. This scale is calibrated in units of "quefrency," i.e., pitch period, in accordance with Bogert, Healy, and Tukey.[1] The desired frequency linearity is achieved by frequency-stepping the spectrum waveform in the storage loop by the same amount on each circulation.

C. Processing Period

Since the spectral-speech characteristics of talkers differ widely, no single value of processing period P of

[8] U. S. Patent No. 3,013,209 and other domestic and foreign patents pending.

[9] M. R. Weiss and C. M. Harris, J. Acoust. Soc. Am. 35, 207–214 (1963).

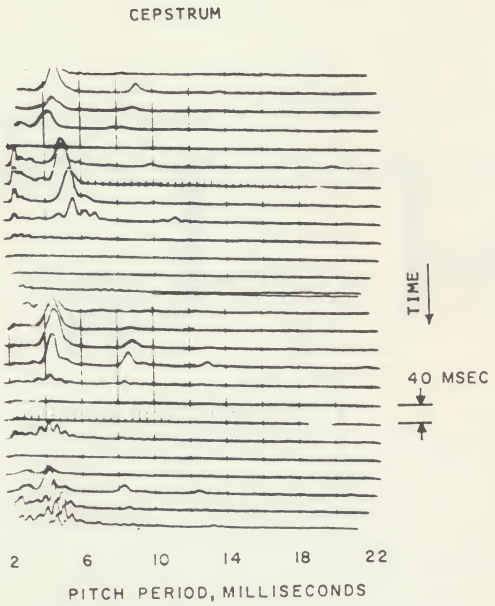

FIG. 3. Cepstrum analyses of the speech of a female talker, recorded from a telephone.

the first analyzer is optimum for all talkers or talking conditions. It is desirable that the processing period be variable in the range between 14 to 48 msec to accomodate at one extreme a maximum rate-of-change of pitch of 3 cps per millisecond, or at the other extreme a minimum pitch of 40 cps. Actually, the processing period can be varied from 2 to 60 msec by use of a selector switch, permitting the extraction of pitch for other signals as well as for speech. The frequency resolution of the analyzer for unweighted input signals is given approximately as $0.85/P$. For example, for a continuous input signal of constant amplitude and fixed frequency, the nominal resolution of the analyzer is 21 cps for a 40-msec processing period and 34 cps for a 25-msec processing period. The speech signal is amplitude-weighted during each processing period. Although other types of weighting functions may be used with the pitch extractor described here, it is convenient to employ the following triangular weighting function $w(t)$:

$$\begin{aligned} w(t) &= t/P, && 0 \leqslant t < P/2, \\ &= 1 - t/P, && P/2 \leqslant t \leqslant P, \\ &= 0, && 0 > t > P. \end{aligned} \tag{1}$$

The effect of this weighting function is to concentrate the energy of a spectral component in the region of its related peak in the output of the spectrum analyzer. While reducing the side lobes in the analyzer's response to a signal component, it broadens the frequency resolution by approximately 40%, resulting in a resolution that is given approximately as $1.2/P$.

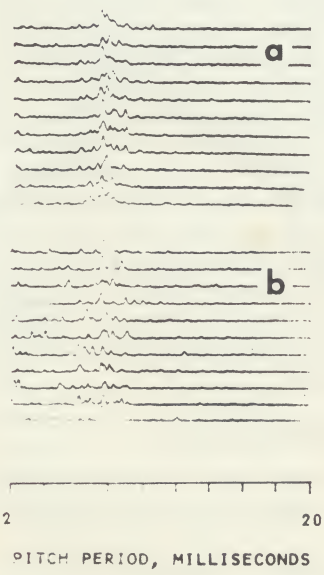

FIG. 4. Comparison of the output of the pitch extractor at a S/N ratio of 0 dB (a) without logarithmic compression and (b) with logarithmic compression.

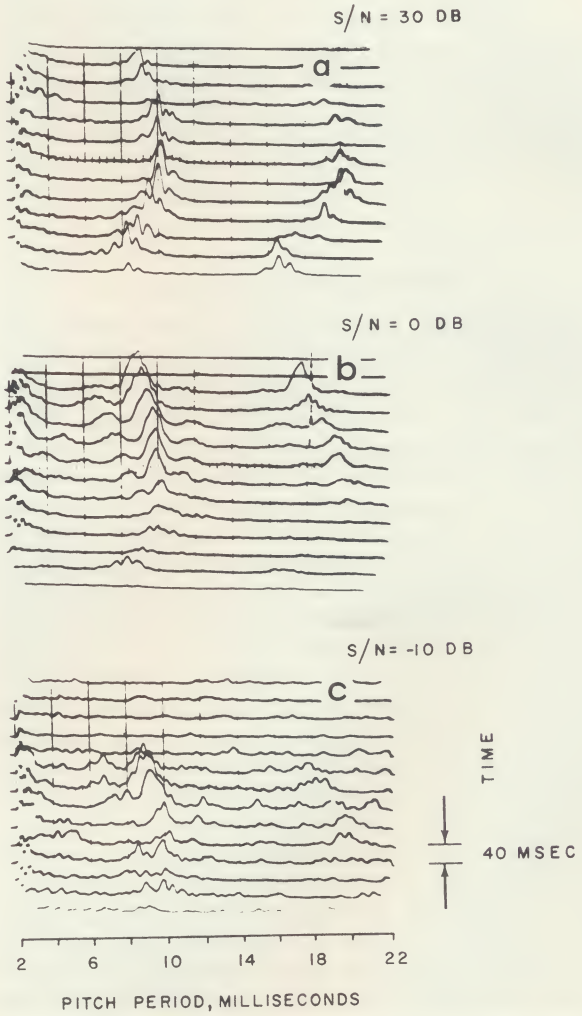

FIG. 5. Output of the pitch extractor for the same input-speech signal (the word *I*, adult male talker) for various S/N ratios (a) S/N=30 dB; (b) S/N=0 to 5 dB; and (c) S/N=−10 to −5 dB.

II. EXAMPLES OF PITCH-EXTRACTOR OUTPUT

This section illustrates the performance of the pitch extractor under practical field conditions. (The examples presented here used a processing period of 40 msec.) The first example, Fig. 3, shows the output of the pitch extractor for a portion of a recorded weather report, spoken by a telephone company operator. Since the telephone channel does not transmit significant components below 300 cps, the fundamental component of the speech was missing. Each major division on the graticule represents a 2-msec interval in pitch period so that the average pitch period indicated by the peaks on the output waveforms is about 4.5 msec, corresponding to a pitch of 222 cps. (Since logarithmic compression was used here, these are "cepstrum" waveforms.) The intervals between voiced speech sounds are characterized by the absence of a major cepstrum peak.

The next example, Fig. 4, compares the response of the pitch extractor under two conditions—with and without logarithmic compression. Here, the S/N ratio was poor: 0 dB. The output without logarithmic compression is slightly "less noisy" than the cepstrum waveform in this case, and energy in the region of the largest peak is more concentrated. The location of the largest peak varies more from trace to trace when logarithmic compression is used.

Figure 5 shows the output of the pitch extractor for the same input-speech signal for various S/N ratios. Figure 5(a) illustrates cepstrum waveforms for an adult male talker for a S/N ratio of approximately 30 dB. The pitch period of the speech signal is 9 msec ($f_0=111$ cps) at the start of the sound, rises to 10.2 msec ($f_0=98$ cps) at the middle, and drops back to 8.4 msec ($f_0=119$ cps) at the end. Next, in Fig. 5(b), noise was added to the speech signal, thereby reducing its S/N ratio to approximately 0 dB. In this case, the spectrum waveform obtained from the first spectrum analyzer was not compressed prior to being applied to the second spectrum analyzer. As a result, the major peaks of the output waveforms are wider and of a somewhat different shape than those shown in Fig. 5(a). However, the locations of these peaks are the same as they were before. Finally, in Fig. 5(c), by the addition of more noise, the S/N ratio was further degraded to approximately −10 dB. Note that the noise has obliterated some of the peaks in the output waveforms. However, major peaks are clearly evident on five of the waveforms. The randomly scattered, low-level peaks that appear in all of the waveforms represent the pitch extractor's response to the noise signal. Even at this poor S/N ratio, a clearly dominant peak in the output of the pitch is evident only when a voiced speech signal is present.

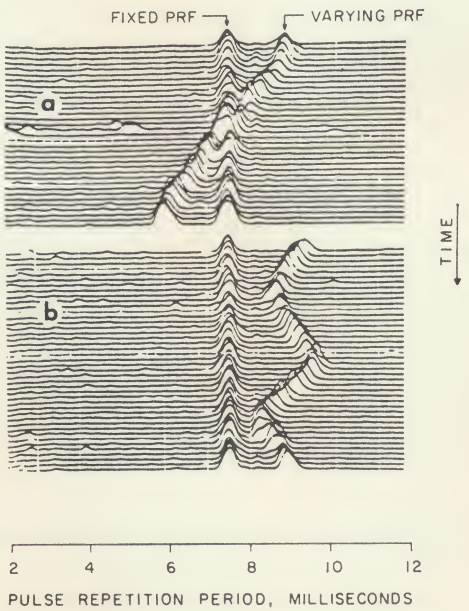

FIG. 6. Double-spectrum analyses of two pulse-train signals that are applied to the input of the pitch extractor simultaneously. One is of fixed pitch frequency. The variable pitch frequency of the second signal (a) crosses that of the first signal, and (b) oscillates near the first signal.

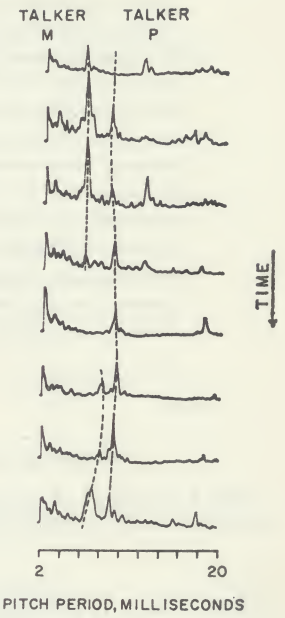

FIG. 7. Double-spectrum analyses of the speech of two adult males talking simultaneously. The pitch track of talker "M" is indicated by the dotted line on the left; the pitch track of talker "P" is indicated by the dotted line on the right.

Consider the double-spectrum analysis of two simultaneously applied input signals. The locations of the peaks in the pitch extractor output will be the same as if the signals were applied individually. An example of the response of the pitch extractor to an input consisting of two pulse trains is shown in Fig. 6. The pulse repetition frequency of one signal was held constant while the other was varied. Note that there is no difficulty in tracking the double-spectrum peaks for each signal. Figure 7 shows the double-spectrum analysis of two speech signals that are applied simultaneously to the analyzer. In this case, the individual double-spectrum peaks are less distinct, and their tracks are less apparent.

Figure 8 shows examples of other applications of the pitch extractor. Figure 8(a) shows the cepstrum wave-

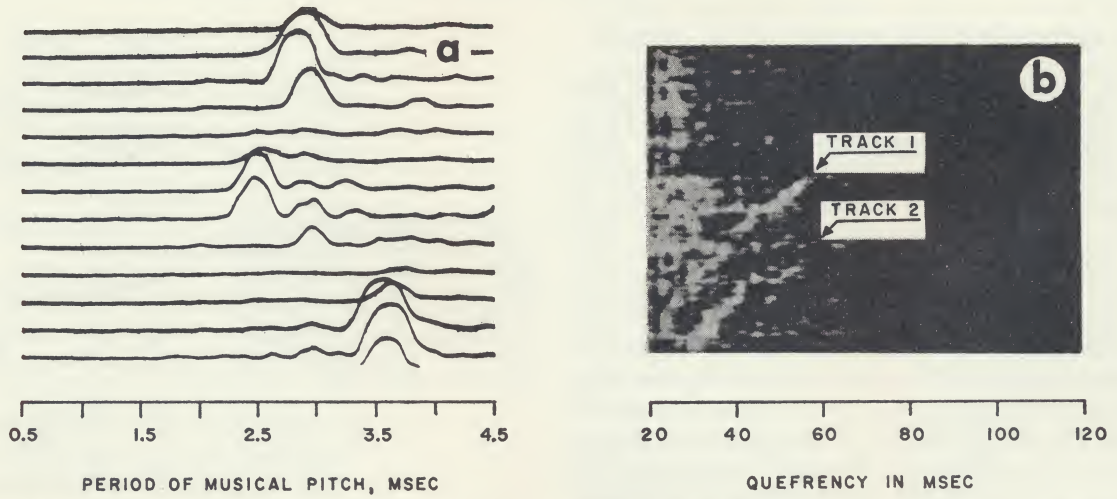

FIG. 8. Examples of the output of the pitch extractor. (a) Cepstrum waveform of sound of a guitar. (b) Double-spectrum waveform of passive sonar signal.

forms of sounds of a guitar. Changes in pitch corresponding to different notes are evident. Figure 8(b) shows an analysis of a passive sonar signal. Here, the double-spectrum waveforms (photographed as intensity-modulated oscilloscope traces) indicate that the sonar signal contained two quasiperiodic signals whose fundamental frequencies changed in a similar manner.

ACKNOWLEDGMENT

This system, based on the double-spectrum analysis technique, was designed and constructed by the Federal Scientific Corporation under the sponsorhsip of the Air Force Cambridge Research Laboratories, Office of Aerospace Research.